**Data Analysis Project**

BIOST311: Regression Methods in the Health Sciences

Taylor Okonek and Charlie Wolock

Spring 2022

**Overview**

The course project is a data analysis, consisting of five parts: (1) a project proposal; (2) a statistical analysis plan; (3) a draft report; (4) a peer review; and (5) a final report. We want you to implement components of the analysis as you learn new concepts in both class and discussion section. Our goal in having multiple parts to the project, due at different points throughout the quarter, is to provide feedback (and have your classmates provide feedback) multiple times so that your final report is as polished as possible. With that in mind, please do the best job you can on each component—the better the document you hand in, the better comments we can give you!

There are multiple valid methods to choose from for this project. You may choose to analyze the data differently from us; spend your time choosing a method that you can justify, implement, and interpret correctly.

Each of the following components of the project has its own guidelines, each listed in separate documents. The project is worth 30% of your overall course grade.

**The project proposal** (due April 15, 5% of the overall course grade)

This component of the project addresses two questions:

* What dataset are you using?
* What scientific questions are you answering?

We encourage you to find a dataset that interests you—since being excited about the data will make you more excited about analyzing it—but we have a few criteria that each dataset must meet.

1. The dataset must have more than five variables

2. The dataset must have an outcome variable that is continuous, binary, or time-to-event

3. You must be able to pose at least two interesting scientific questions (judged by us) that you can answer using the data

Some of you may already be involved in research—If you can get permission to use these data, we highly encourage it!

If you don't have a research project, a couple of options for finding datasets are

* [University of California at Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php)
* [Kaggle competition datasets](https://www.kaggle.com/datasets)
* [FiveThirtyEight](https://data.fivethirtyeight.com/)

We also have a handful of datasets that you can use if you're unable to find one that interests you. Please ask both of us if this is the case, and we will provide a dataset that you (hopefully) find interesting!

We will provide you with an R Markdown template for this component of the project.

**The statistical analysis plan** (due April 29, 5% of the overall course grade)

In a statistical analysis plan (SAP), you write down explicitly all the statistical analyses (e.g., descriptive statistics, plots and tables, and inferential analyses) that you will conduct using the data. See the guidelines for the SAP for more information about the required elements.

An additional component of the statistical analysis plan is to run one preliminary analysis on your data, using one of the techniques we have learned in class or discussion section.

We will provide you with an R Markdown template for this component of the project.

**The first draft** (due May 20, 5% of the overall course grade)

The final report is due on June 8, but we think you should get started early, and we want you to have a high-quality report by the end of the quarter. With that in mind, we're asking that you hand in a first draft with time for us to read your drafts and hand back comments, and for your classmates to complete the peer review.

See the report document for more information. We will not provide you with any templates for this component of the project.

**Peer review** (due May 27, 5% of the overall course grade)

You will be paired with one of your classmates to complete a peer review of the first draft. You will read your partner’s draft and provide comments and critiques regarding the statistical analysis, writing, and overall effectiveness.

See the peer review document for more information. We will provide you with an R Markdown template for this component of the project.

**The final report** (due June 8, 10% of the overall project grade)

You will write up a full account of your statistical analysis in this report, including: your scientific questions; your statistical hypotheses; your statistical methods, including for descriptive statistics, figures and tables, and inferential analyses; your results; and your statistical and scientific conclusions. All of this will be written in a style suitable to be included in a manuscript.

By turning in a draft earlier in the quarter, we hope that you will have enough time to incorporate feedback before handing in the final version.

We will provide guidelines on the components of the report in a separate document but will not provide you with a template file.